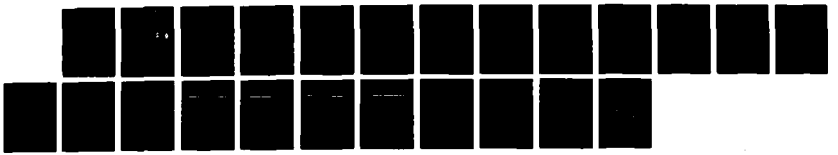
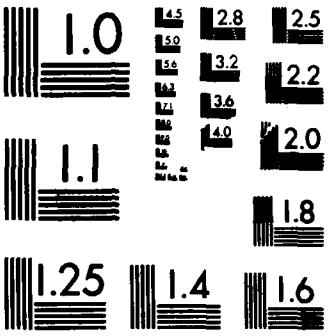
AD-A183 064 KEYWORD CLUSTER ALGORITHM FOR EXPERT SYSTEM RULE BASES 1/1
(U) AEROSPACE CORP EL SEGUNDO CA COMPUTER SCIENCE LAB
S LINDELL 22 JUN 87 TR-0086A(2920-02)-1 SD-TR-87-36
UNCLASSIFIED F04701-85-C-0086 F/G 12/9 NL





REPORT SD-TR-87-36

Keyword Cluster Algorithm for Expert System Rule Bases

S. LINDELL
Computer Science Laboratory
Laboratory Operations
The Aerospace Corporation
El Segundo, CA 90245

22 June 1987

Prepared for

SPACE DIVISION
AIR FORCE SYSTEMS COMMAND
Los Angeles Air Force Station
P.O. Box 92960, Worldway Postal Center
Los Angeles, CA 90009-2960

This report was submitted by The Aerospace Corporation, El Segundo, CA 90245, under Contract No. F04701-85-C-0086 with the Space Division, P.O. Box 92960, Worldway Postal Center, Los Angeles, CA 90009-2960. It was reviewed and approved for The Aerospace Corporation by H. R. Rugge, Acting Director, Computer Science Laboratory.

Lt David Rosenberg/CLVA was the project officer.

This report has been reviewed by the Public Affairs Office (PAS) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication. Publication of this report does not constitute Air Force approval of the report's findings or conclusions. It is published only for the exchange and stimulation of ideas.

DAVID ROSENBERG, Lt, USAF
MOIE Project Officer
SD/CLVA

JOSEPH HESS, GM-15
Director, AFSTC West Coast Office
AFSTC/WCO OL-AB

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER SD-TR-87-36	2. GOVT ACCESSION NO. A183 064	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) KEYWORD CLUSTER ALGORITHM FOR EXPERT SYSTEM RULE BASES		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER TR-0086A(2920-02)-1
7. AUTHOR(s) Suzanne Lindell		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS The Aerospace Corporation El Segundo, CA 90245		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Space Division Air Force Systems Command Los Angeles CA 90009		12. REPORT DATE 22 June 1987 13. NUMBER OF PAGES 18
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Assertion Cluster Graph
Keyword Cluster Algorithm
Expert Systems
Display Tool
Rule Based Systems
Knowledge-Based Systems
Knowledge Engineering

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

An algorithm is described for automatically organizing a fairly unstructured expert system rule base in order to facilitate updating and debugging by programmers. The algorithm operates on a structure called an Assertion Cluster Graph (ACG) which consists of nodes for every assertation in the rule base and of arcs connecting the assertations that are dependent on each other for their values. The algorithm reduces the complexity of ACG by replacing related groups of assertions in the graph by a single summary node. The —

DD FORM 1473 (FACSIMILE)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

19. KEY WORDS (Continued)

20. ABSTRACT (Continued)

assertions are clustered into groups according to Keywords contained in their English Text. The algorithm is used to create an interactive program which displays the summarized version of the ACG and can expand the clusters on command. It is anticipated that this expert system display tool will not only be helpful to programmers, but will also enable users to better understand how the system works.

CONTENTS

FIGURES

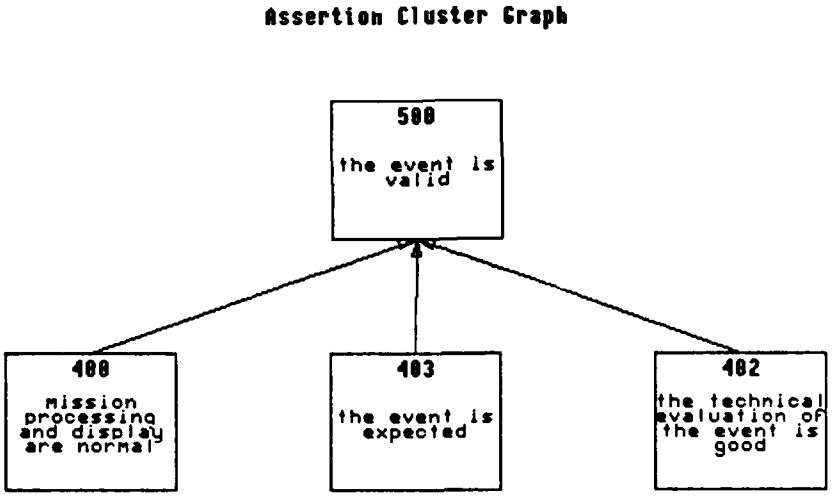
I. INTRODUCTION

The keyword cluster algorithm is part of a project to create tools for building and organizing expert systems. The goal of the project is to find intelligent ways of organizing a fairly unstructured expert system rule base, in such a way as to facilitate updating and debugging of the rule base by its programmers. The organized rule base should also be easier for the users of the expert system to understand. Other facets of this project are being developed by Kirstie Bellman, April Gillam, Paul Mazaika, and Rod McGuire.

Other expert system building tools, such as EMYCIN and EXPERT, are intended to be used from the beginning of the system design process, whereas the keyword cluster algorithm attempts to structure an already established rule base built without the aid of any tools. It produces an overview of the expert system in which its workings are more clearly evident than in the original collection of rules.

II. ASSERTION CLUSTER GRAPH

Many expert systems consist of production rules of the form IF some facts are currently asserted to be true (or false), THEN assert some other fact to be true (or false). The facts in the antecedents and the consequents of the rules are called assertions. Since the consequent of each rule is part of the antecedent of some other rule, the connections between the assertions can be displayed as a tree-like graph called an assertion cluster graph (or ACG). In the ACG, each assertion is represented by a node, and there is a directed arc connecting two nodes if there is a rule (or rules) in which the assertion at the tail of the arc is one of the antecedents of the rule, and the assertion at the head of the arc is the consequent. (See Figure 2-1). The node at the tail of the arc is called the child of the node at the head. The information of how the assertions combine to produce the consequent is omitted in this graph, but it does show the dependencies between assertions.



Rule 1: IF 400=TH and 402=FH and 403=TH THEN 500=FH
Rule 2: IF 400=TH and 402=TH and 403=FH THEN 500=TM

Figure 2-1. Example ACG and Related Rules

III. ALGORITHM DESCRIPTION

A. GENERAL IDEA

The purpose of the keyword cluster algorithm is to takc the ACG, which although it is smaller than the entire rule base, is still rather large and cumbersome, and to condense it by grouping some related assertions into clusters. In the condensed ACG, those assertions are replaced by a node representing the cluster. The method of clustering assertions is according to topic, i. e., if a number of assertions have the same topic, then they are put into the same cluster. The cluster is named according to the topic of the assertions that it contains. In the rule base upon which the work in this paper is based, the only information as to the topic of an assertion is the assertion text supplied by the domain expert. Therefore, if the texts of two assertions have some meaningful words in common, they should be in the same cluster. Those common words are called keywords, and they are chosen to describe the cluster.

B. KEYWORD SELECTION

The keyword finding algorithm is designed to select the most specific and meaningful keywords as is possible using the method of comparing assertion texts.

1. MEANINGFULNESS CRITERIA

Words are considered meaningful if they are not common parts of speech such as articles, prepositions, or connectives. Words that are so common to the assertions that they appear in a lot of them regardless of where they are in the ACG are not considered meaningful, and so are filtered out (not automatically).

2. SPECIFICNESS CRITERIA

In some cases, the keywords found by comparing two assertions may be contained by those found by comparing two others. To generate the most specific keywords, that set of keywords with the greatest number of words is chosen, and the smaller sets are rejected. For example, in the sample rule base upon which the algorithm was performed, which has to do with classifying detected events, there are a number of assertions having to do with blanks. Some of those blanks are moving and some are static. The words common to two assertions about static blanks would be "static blank", whereas the word common to an assertion about static blanks and an assertion about moving blanks would simply be "blank". The words "static blank" are more specific than "blank", so they would be preferred by the algorithm. Thus a cluster called "static blank" would be created, rather than one called "blank".

C. CLUSTER TYPES

The keyword cluster algorithm selects nodes with a lot of children, and puts those children and their descendents into three kinds of clusters, "major clusters", "regular clusters", and "subclusters":

a. Major clusters consist of those children of a selected node that have similar topics, and all of their respective descendents.

b. Regular clusters consist of one child and all of its descendents. It is assumed that the topic of the descendents of the child is the same as that of the child. Ideally, each child of a selected node should be part of a regular cluster.

c. Subclusters are regular clusters in which all of the assertions are already part of a major cluster.

D. CLUSTERING EXAMPLE

Suppose there is a expert system containing the rule, IF Mr. X owns a new red motor car, or Mr. X owns a red motor scooter, or Mr. X owns a red bicycle, or Mr. X owns ice skates, or Mr. X owns roller skates, THEN Mr. X owns a form of transportation. The assertion in the consequent of the above rule, Mr. X owns a form of transportation, has five children.

1. MAJOR CLUSTER CONSTRUCTION

Two of the children describe motorized vehicles owned by Mr. X, and their texts have the word "motor" in common. So those children and their descendents can form the major cluster named MOTOR. Two other children describe some kind of skates owned by Mr. X, and they can likewise form the major cluster SKATES.

2. REGULAR AND SUBCLUSTER CONSTRUCTION

Suppose the assertion Mr. X owns a new red motor car has the following children:

a. Mr. X drives a car to work.

b. Mr. X took out a new car loan last month.

c. Mr. X is seen poslishing a red car in his driveway every Saturday.

Then the assertion Mr. X owns a new red motor car and its children can form the cluster CAR. Likewise, the other children of Mr. X owns a form of transportation can form the clusters MOTOR SCOOTER, BICYCLE, ICE SKATES, and ROLLER SKATES. The clusters CAR and MOTOR SCOOTER turn out to be subclusters of the major cluster MOTOR, and the clusters ICE SKATES and ROLLER SKATES are subclusters of the major cluster SKATES.

3. FIGURE

Figure 3-1 shows the condensed ACG for this example. Notice that the clusters MOTOR and SKATES create two extra nodes, but they serve to spread the clusters onto two levels instead of crowding them onto one.

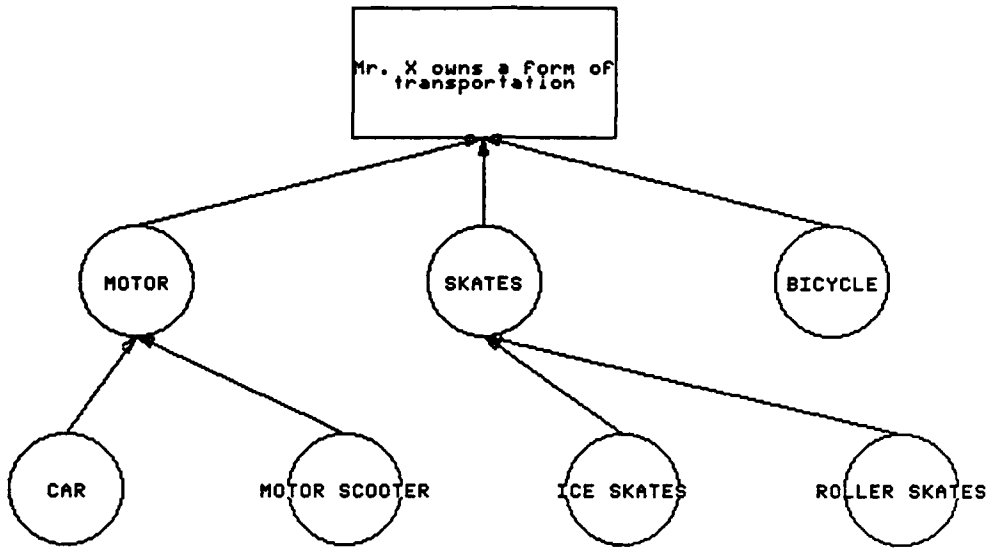


Figure 3-1. Reduced ACG for Transportation Example

IV. ALGORITHM DETAILS

The algorithm itself consists of two parts. The first part of the algorithm searches for major clusters, i.e. those containing sibling assertions with common topics, while the second part determines the keywords for each child assertion's cluster (or subcluster) based on the specific topic of each child and its descendents.

A. PART ONE

1. MATCHING PROCEDURE

To find the major clusters underneath a selected node, all of the children of that node are compared for common words. If a match is discovered between the texts of two assertions, those assertions are temporarily put into a cluster with those keywords as its name. If, later on, one (or both) of those assertions makes a better match, i. e., it is put into a cluster whose keywords contain the keywords for the assertion's old cluster, then it is taken out of the old cluster. If the new keywords don't contain the old ones, (and vice-versa), then the assertion remains in both clusters. If a cluster eventually has one or no assertions in it, then it is thrown out. An assertion can join an already existing cluster by matching one of the assertions in it. Any assertions that don't match other assertions are not put into any cluster. The clusters left at the end are the major clusters, some of which may share assertions.

2. TRANSPORTATION RULE BASE EXAMPLE

In the transportation rule base, the assertion Mr. X owns a new red motor car matches the assertion Mr. X owns a red bicycle, and the common word is "red". (The words "Mr. X owns" are filtered out because they are too common in the rule base, and the "a" is filtered out because it is an article). The two assertions are temporarily put into a cluster named "red". However, the assertion Mr. X owns a new red motor car makes a better match with the assertion Mr. X owns a red motor scooter, because the words "red motor" contain the word "red". So the latter two assertions are put into a cluster, and the assertion Mr. X owns a new red motor car is taken out of the old cluster. The potential major cluster "red" only contains one assertion and is thrown away.

B. PART TWO

It is already known that each child is supposed to be clustered with its descendents. Hence, the purpose of this part of the algorithm is just to generate the keywords for each child's cluster.

1. KEYWORD GENERATION PROCEDURE

To determine the cluster (or subcluster) keywords for a child of a selected node, the child is compared with its descendents but the descendents are not compared with each other. The keyword clusters are chosen by finding the best matches as in part one, and the resulting clusters provide names from which to choose the child's final cluster name.

2. CLUSTER NAME SELECTION

If the child ends up in more than one cluster, then there is a problem of what keywords to use as the cluster name. The algorithm attempts to create a new name by looking for common words among the different cluster names. If it finds some, it selects those as the keywords for the child's cluster. For example, in the transportation rule base, the algorithm would create two clusters for the child assertion Mr. X owns a new red motor car, the first with the keywords "new car", and the second with the keywords "red car". The two cluster names share the word "car", which would be selected as the cluster name.

3. POOR SUBCLUSTER NAME ELIMINATION

The cluster created for each child in the second part of the algorithm may turn out to be a subcluster of some major cluster. In that case, there is another step in the algorithm.

a. Procedure

For each child already in a major cluster, the subcluster name(s) is (are) compared with the name of the major cluster. If the keywords for a subcluster are contained by the major cluster's keywords, the former keywords are thrown out. This is done because the subcluster's name should be more specific than the major cluster's name, not less.

b. Example

In the sample rule base, a major cluster with the keywords "SPECULAR REFLECTIONS" was created in the first part of the algorithm. The text of one of the assertions (no. 442) in that cluster is "Specular reflections off water is a likely clutter source". In the second part of the algorithm, the cluster name generated for 442 and its descendents was "SPECULAR". Since the phrase "SPECULAR" is contained by the phrase "SPECULAR REFLECTIONS", it is less specific, and was thrown out by the algorithm.

4. EXCEPTION

There is a minor exception to the previous algorithm description. If a child has a lot of its own children, it is not included in the second part of the algorithm, i. e., it is not clustered with its descendents. The reason for this exclusion is that its own children will be put into clusters themselves by the algorithm, and some extra searching will be eliminated.

V. INTERACTIVE DISPLAY

A. CONDENSED ACG DISPLAY

This algorithm was put into use to create an interactive display of the condensed ACG with keyword clusters. In the display, each node in the ACG is represented by a box containing the number of the corresponding assertion and its text. (See figure 5-1). Under each node whose children (and their descendents) are replaced by clusters, there are circles (one for each cluster) containing the cluster's keywords. In Figure 5-1, the assertions numbered 422 and 127 have keyword clusters underneath them. The clusters for each child (or the major clusters for a group of children) are on the first level, and the subclusters (if any) are on the second level. There are two major clusters for the children of assertion 422, "CLUTTER" and "BACKGROUND", and one regular cluster "SUN". The cluster "BACKGROUND" has a subcluster "BACKGROUND AREA". The category "CLUTTER" has no subclusters. The children of assertion 127 are combined into two major clusters, "POINTS" and "ONE CLOSE STATIC BLANK". Most of the children have clusters underneath them, so there are no subclusters.

B. KEYWORD EXPANSION DISPLAY

Each circle can be expanded to show the subtree of assertions (nodes) contained in that cluster by the click of a mouse button in the circle. Figure 5-2 shows the subtree of nodes generated when the keyword "CLUTTER" is expanded. There are two assertions in the cluster "CLUTTER", nos. 460 and 435. Assertion 460's children are in clusters themselves.

C. SUBTREE DISPLAY

The subtree (all of the children and their descendents/clusters) rooted at a node with clusters under it can be displayed all at once by the click of a mouse button in the node's box. Figure 5-3 is the subtree rooted at assertion 422. It exposes some of assertion 422's children which could not be put in any cluster in Figure 5-1, namely, assertions 426, 429, 436, and 445. Figure 5-4 is the subtree rooted at assertion 127. Assertions 144, 145, and 146 all have the same children, so that is why there are three arcs coming from two of the clusters. The four clusters on the first level are all major clusters. The cluster "STATIC BLANK" contains two assertions, each of which is in its own subcluster, "SOLAR BLANK" and "LUNAR BLANK". For those assertions dealing with static blanks, the algorithm worked particularly well.

D. RULE DISPLAY

The display also has the capability to list the rules associated with any two assertions, since they are not explicit in the ACG itself. The two assertions are selected by pushing a mouse button down in one assertion's box and lifting it up in the other. Figure 5-5 shows the same ACG as in Fig. 5-1, with an extra window to display the rule texts. The rules shown are those whose antecedent is assertion 410, and whose consequent is assertion 500.

E. INTENT OF DISPLAY TOOL

This display tool is not an actual rule base building tool, since there is no provision to add or delete rules. It is mainly intended for the expert system users. However, the capability of modifying the rule base could easily be added.

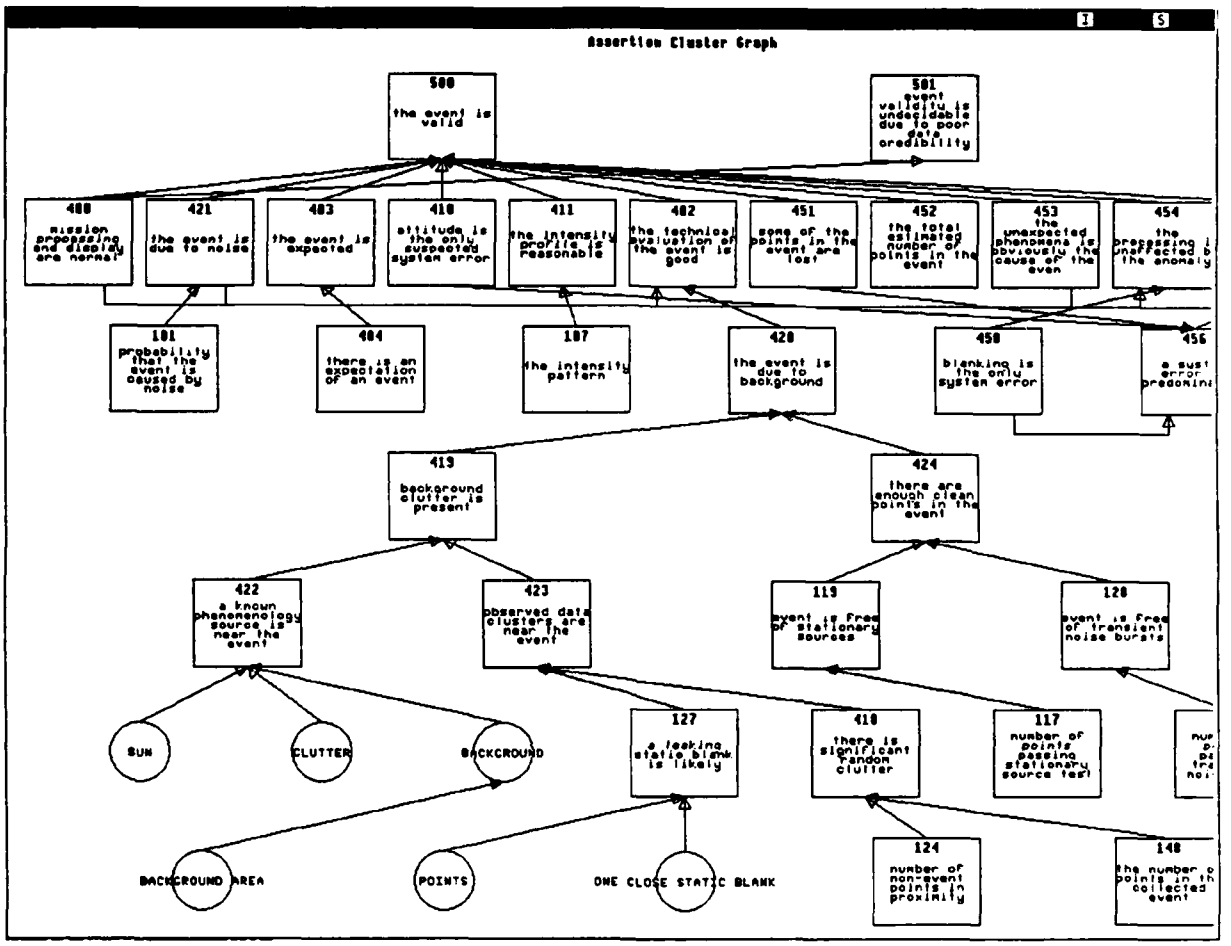


Figure 5-1. Condensed ACG with Keyword Clusters, Main Display

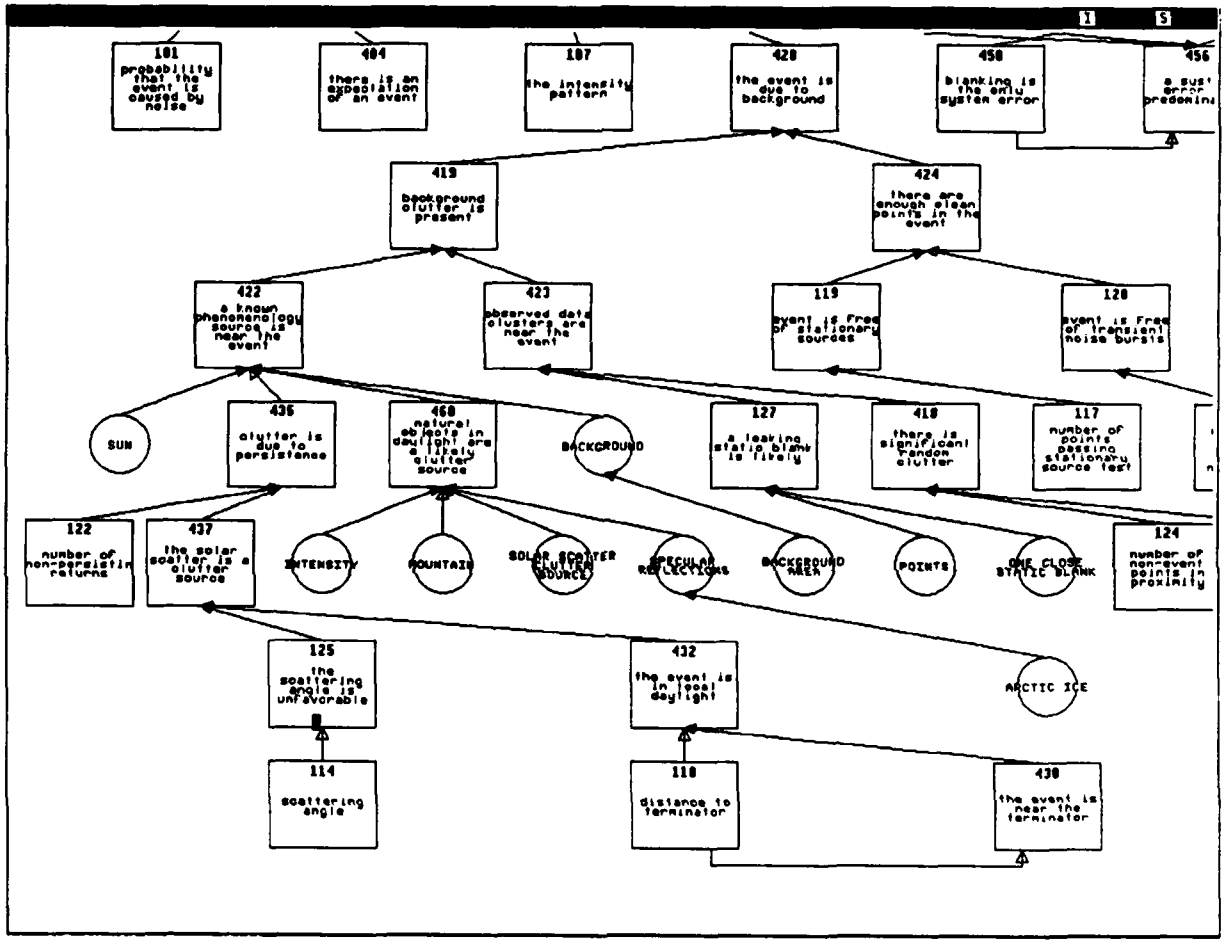


Figure 5-2. Keyword "CLUTTER" Expanded

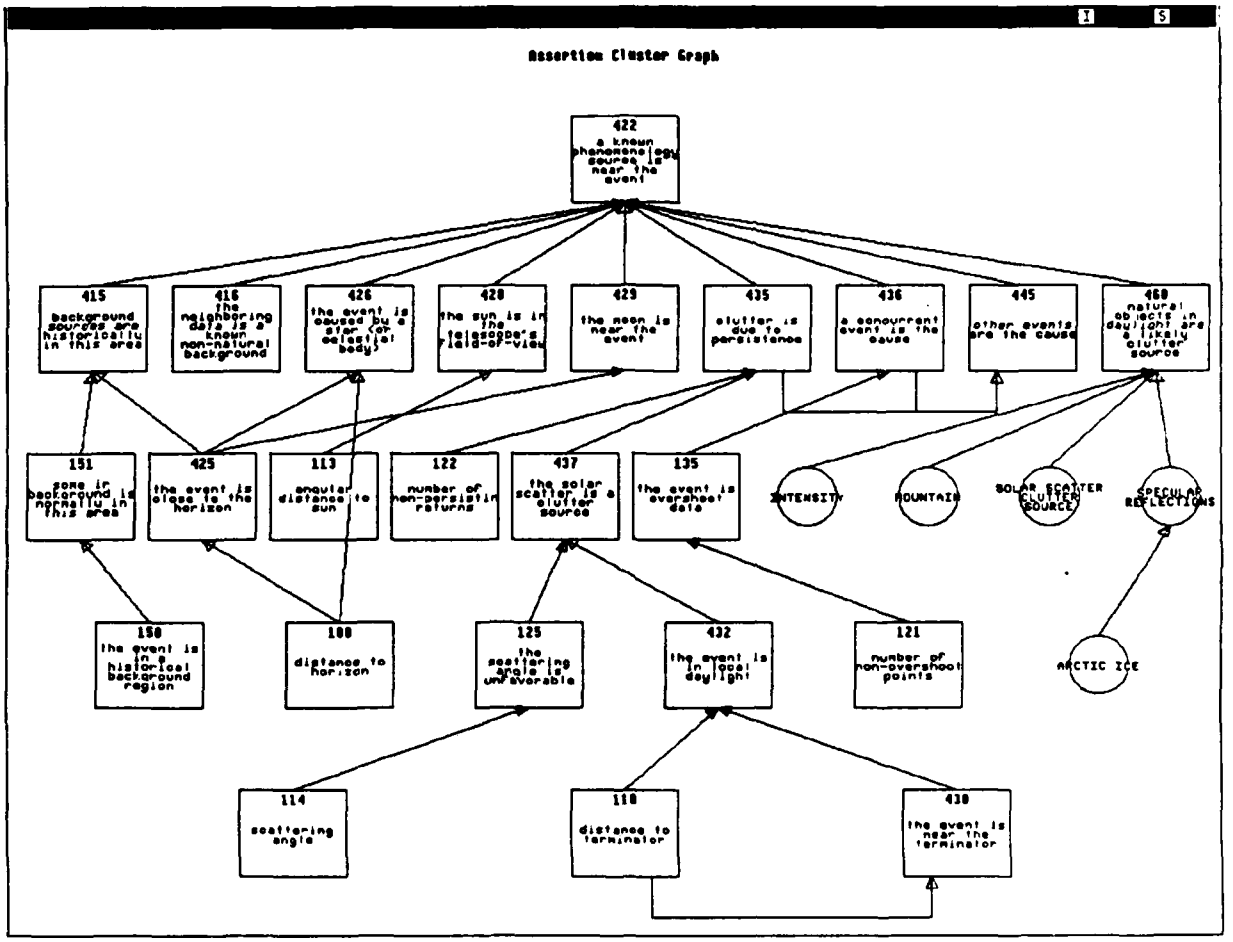


Figure 5-3. Subtree Rooted at Assertion 422

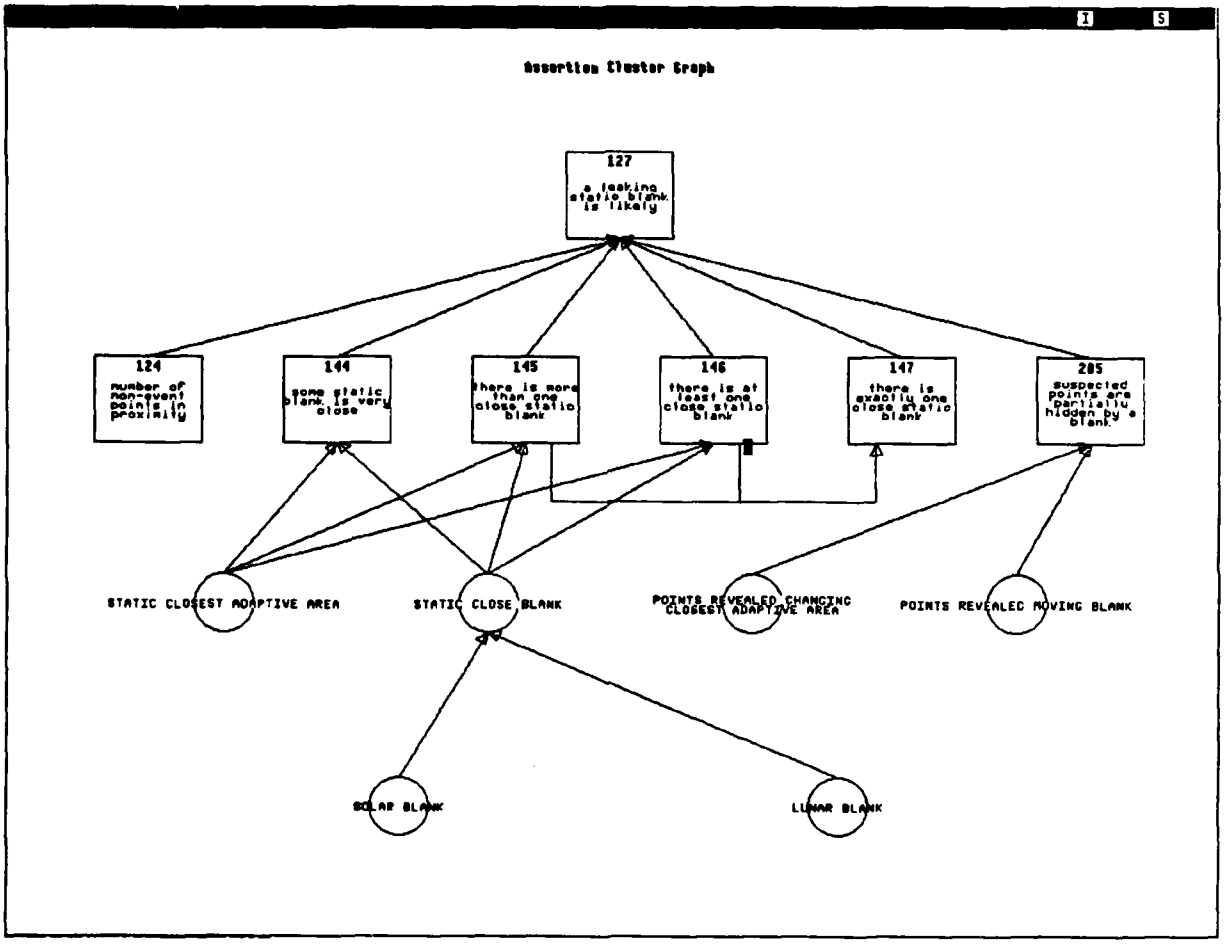


Figure 5-4. Subtree Rooted at Assertion 127

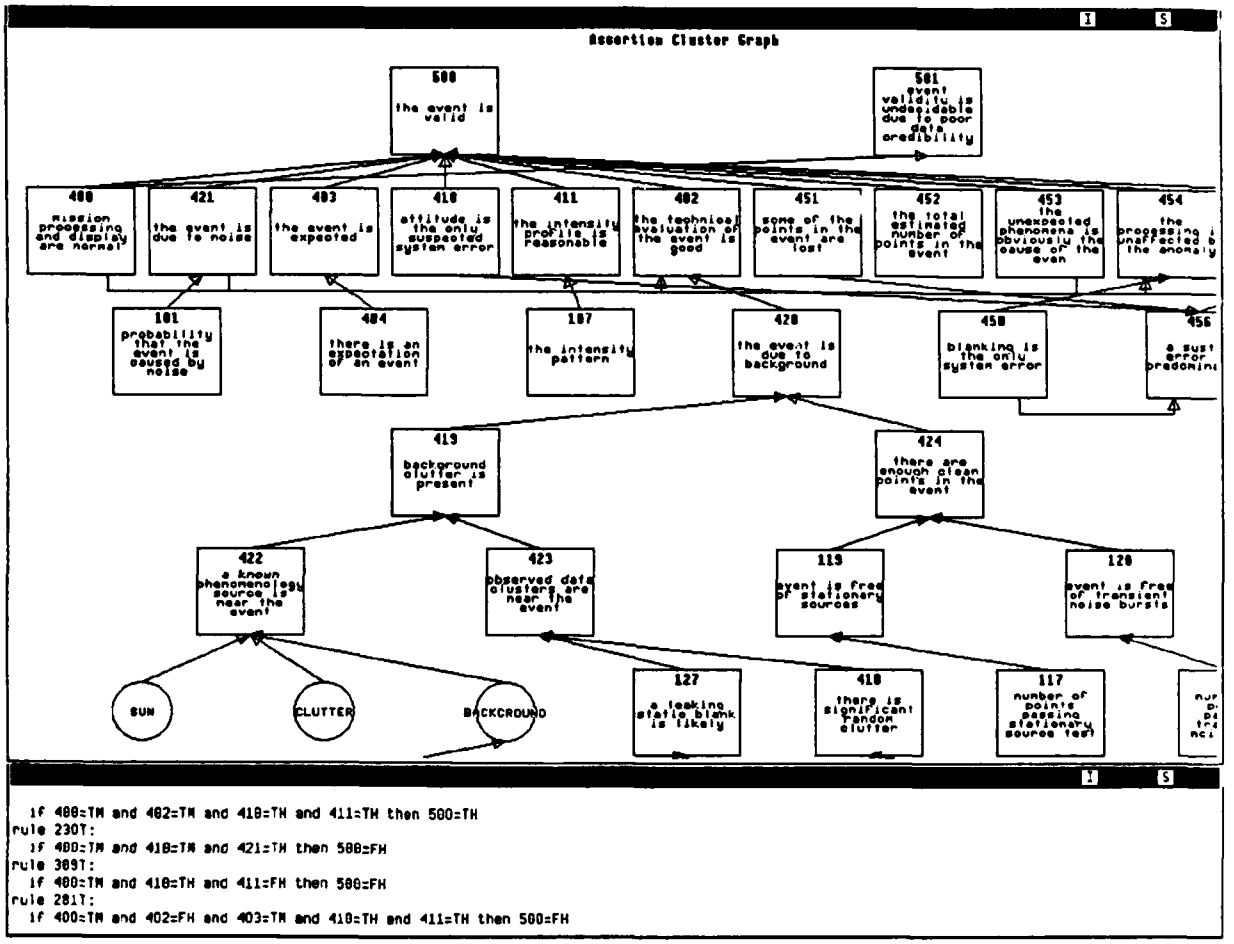


Figure 5-5. Rules Associated with Assertions 410 and 500

VI. LIMITATIONS

The method of choosing keywords by looking for common phrases in the assertions texts reduces the apparent complexity of the ACG, but does not necessarily produce meaningful keywords, or, in fact, any at all.

A. MEANINGFULNESS CRITERIA

The criteria for deciding meaningfulness of phrases needs to be more sophisticated. For instance, in the above examples, the cluster "POINTS" isn't very specific, and the subcluster "BACKGROUND AREA", isn't much different from its major cluster "BACKGROUND". The clusters "POINTS REVEALED MOVING BLANK" and "SOLAR SCATTER CLUTTER SOURCE" seem to have too many words; "MOVING BLANK" and "SOLAR SCATTER" would have been better choices.

B. SEMANTIC CONSIDERATION

In comparing words, the semantics of the words needs to be taken into account, as well as just the syntax. For example, the most important keyword in the text of assertion 442 is "water". None of its descendents have the word "water" in their texts, but some of them have the word "sea". A semantic-based keyword search would consider the two words similar, and pick one of them as the cluster name. Ideally, the clusters should be generated by just looking at the text of the assertions and picking out the most important phrase, and by not comparing them to each other. That, however, seems to be a problem on the forefront of artificial intelligence research.

VII. CONCLUSION

The keyword cluster algorithm is a method for automatically organizing and summarizing large expert system rule bases by matching phrases in the English text of the assertions. The algorithm was partially successful in organizing the rule base on which it was tried. It reduced the complexity of the ACG, and generated somewhat meaningful keywords. Its main limitation is that it does not use any semantic knowledge of the text. However, despite this limitation, it is helpful in structuring the rule base.

LABORATORY OPERATIONS

The Aerospace Corporation functions as an "architect-engineer" for national security projects, specializing in advanced military space systems. Providing research support, the corporation's Laboratory Operations conducts experimental and theoretical investigations that focus on the application of scientific and technical advances to such systems. Vital to the success of these investigations is the technical staff's wide-ranging expertise and its ability to stay current with new developments. This expertise is enhanced by a research program aimed at dealing with the many problems associated with rapidly evolving space systems. Contributing their capabilities to the research effort are these individual laboratories:

Aerophysics Laboratory: Launch vehicle and reentry fluid mechanics, heat transfer and flight dynamics; chemical and electric propulsion, propellant chemistry, chemical dynamics, environmental chemistry, trace detection; spacecraft structural mechanics, contamination, thermal and structural control; high temperature thermomechanics, gas kinetics and radiation; cw and pulsed chemical and excimer laser development including chemical kinetics, spectroscopy, optical resonators, beam control, atmospheric propagation, laser effects and countermeasures.

Chemistry and Physics Laboratory: Atmospheric chemical reactions, atmospheric optics, light scattering, state-specific chemical reactions and radiative signatures of missile plumes, sensor out-of-field-of-view rejection, applied laser spectroscopy, laser chemistry, laser optoelectronics, solar cell physics, battery electrochemistry, space vacuum and radiation effects on materials, lubrication and surface phenomena, thermionic emission, photosensitive materials and detectors, atomic frequency standards, and environmental chemistry.

Computer Science Laboratory: Program verification, program translation, performance-sensitive system design, distributed architectures for spaceborne computers, fault-tolerant computer systems, artificial intelligence, microelectronics applications, communication protocols, and computer security.

Electronics Research Laboratory: Microelectronics, solid-state device physics, compound semiconductors, radiation hardening; electro-optics, quantum electronics, solid-state lasers, optical propagation and communications; microwave semiconductor devices, microwave/millimeter wave measurements, diagnostics and radiometry, microwave/millimeter wave thermionic devices; atomic time and frequency standards; antennas, rf systems, electromagnetic propagation phenomena, space communication systems.

Materials Sciences Laboratory: Development of new materials: metals, alloys, ceramics, polymers and their composites, and new forms of carbon; nondestructive evaluation, component failure analysis and reliability; fracture mechanics and stress corrosion; analysis and evaluation of materials at cryogenic and elevated temperatures as well as in space and enemy-induced environments.

Space Sciences Laboratory: Magnetospheric, auroral and cosmic ray physics, wave-particle interactions, magnetospheric plasma waves; atmospheric and ionospheric physics, density and composition of the upper atmosphere, remote sensing using atmospheric radiation; solar physics, infrared astronomy, infrared signature analysis; effects of solar activity, magnetic storms and nuclear explosions on the earth's atmosphere, ionosphere and magnetosphere; effects of electromagnetic and particulate radiations on space systems; space instrumentation.

...

END

9-87

DTIC